Agent Evaluation Metrics — Comprehensive One■Page Cheat Sheet

Retrieval & Discovery

Metric	What it Measures
Recall / Recall@K	Coverage of all relevant items; missing evidence detection
Precision / Precision@K	Noise level in retrieved results
Hit Rate	Whether at least one relevant item was retrieved
Coverage	Breadth of topics/domains surfaced

Ranking Quality

Metric	What it Measures
MRR	How early the first relevant result appears
NDCG	Overall ranking quality with graded relevance
MAP	Mean precision across ranking positions

Reasoning & Task Success

Metric	What it Measures
Exact Match	Perfect correctness of answer
F1 Score	Partial correctness of structured outputs
Task Success Rate	End■to■end completion of objectives
Utility Score	Practical usefulness of result

Faithfulness & Grounding

Metric	What it Measures
Citation Accuracy	Claims supported by cited sources
Groundedness	Output traceable to retrieved evidence
Hallucination Rate	Unsupported or fabricated statements

Efficiency & Cost

Metric	What it Measures
Latency	Time to complete task
Token Usage	LLM cost footprint
Tool Calls	Efficiency of action selection
Planning Steps	Over■planning or looping behavior

Human Evaluation

Metric	What it Measures
Relevance	Alignment with user intent
Coherence	Logical organization of output
Trustworthiness	User confidence in result
Preference	A/B comparison quality

Rule of thumb: agent failures are usually retrieval or ranking failures before reasoning failures.